REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
SEPTEMBER 2009

2. REPORT TYPE
Conference Paper Postprint

3. DATES COVERED *(From - To)*
September 2009

4. TITLE AND SUBTITLE
SPEAKER RECOGNITION ON LOSSY COMPRESSED SPEECH USING THE SPEEX CODEC

5a. CONTRACT NUMBER
FA8750-05-C-0029

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
35885G

6. AUTHOR(S)
A.D. Lawson, A.R. Stauffer

5d. PROJECT NUMBER
1049

5e. TASK NUMBER
05

5f. WORK UNIT NUMBER
AB

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research Associates for Defense Conversion, Inc.
10002 Hillside Terrace
Marcy, NY 13403-2102

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/RIEC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2009-45

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited. PA#88ABW-2009-1156 Date Cleared: 25-March-2009

13. SUPPLEMENTARY NOTES
© 2009 ISCA. This paper was published in the Proceedings of the Interspeech 2009 Brighton United Kingdom, 6-10 September-2009. This work is copyrighted. This work was funded in whole or in part by Department of the Air Force contract number FA8750-05-C -0029. The U.S. Government has for itself and others acting on its behalf an unlimited, paid-up nonexclusive, irrevocable worldwide license to use, modify, reproduce, release, perform, display, or disclose the work by or on behalf of the Government. All other rights are reserved by the copyright owner.

14. ABSTRACT
This paper examines the impact of lossy speech coding with Speex on GMM-UBM speaker recognition (SR). Audio from 120 speakers was compressed with Speex into twelve data sets, each with a different level of compression quality from 0 (most compressed) to 10 (least), plus uncompressed. Experiments looked at performance under matched and mismatched compression conditions, using models conditioned for the coded environment, and Speex coding applied to improving SR performance on other coders. Results show that Speex is effective for compression of data used in SR and that Speex coding can improve performance on data compressed by the GSM codec.

15. SUBJECT TERMS
Speaker identification, speech coding

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	UU	5

19a. NAME OF RESPONSIBLE PERSON
Michelle Grieco

19b. TELEPHONE NUMBER *(Include area code)*
N/A

Speaker Recognition on Lossy Compressed Speech using the Speex Codec

A. R. Stauffer, A. D. Lawson

RADC, Inc., Rome, NY USA

stauffar@clarkson.edu, Aaron.Lawson.ctr@rl.af.mil

Abstract

This paper examines the impact of lossy speech coding with Speex on GMM-UBM speaker recognition (SR). Audio from 120 speakers was compressed with Speex into twelve data sets, each with a different level of compression quality from 0 (most compressed) to 10 (least), plus uncompressed. Experiments looked at performance under matched and mismatched compression conditions, using models conditioned for the coded environment, and Speex coding applied to improving SR performance on other coders. Results show that Speex is effective for compression of data used in SR and that Speex coding can improve performance on data compressed by the GSM codec.

Index Terms: speaker identification, speech coding

1. Introduction

Compressed audio has come to play an enormous role in modern communications infrastructure. VOIP, voicemail, telephony, archival audio storage, internet streaming audio, and real time gaming communications have all come to use lossy speech coding as the primary method of compressing audio for storage or transmission. Despite the ubiquity of lossy compression, speaker recognition research has been sparse and mainly examined at the effect of GSM coding [1] [2]. Little or no studies have dealt with newer free coders, such as Speex [3] [4] or Vorbis [5]. The direct impact of high rates of lossy compression, compression as a source of channel mismatch, and on the use of compressed and composite models from various coding techniques is largely unexplored. This is in contrast to the hundreds of studies that have undertaken on the impact of telephone handsets, live microphones, channel mismatch, and noise. Considering the prevalence of podcasts, streaming internet communications, telephony and VOIP it is clear that lossy compressed audio transmission makes up an enormous part of human communication. Speaker recognition in the compressed environment is thus clearly of interest to many, including the communications industry, gaming industry, forensics/law enforcement, and those involved in speaker biometrics and verification.

1.1. Goals of this study

The aim of this study is thus to quantify the impact of lossy speech coding on speaker recognition by applying the state of the art, freely available and speech specific Speex compression algorithm [3] to audio files in SR trials. Our major research questions are: 1) to what extent does coding impact SR when compression conditions are matched? 2) what is the impact of compression mismatch or using uncompressed models on coded speech? 3) What levels of compression, if any, are possible without significantly impacting recognition? 4) Can a model set be conditioned to perform across a range of compressed and uncompressed speech? 5) can Speex coding be used to improve SR on other speech coding techniques, such as the common GSM or Vorbis algorithms?

2. Lossy/Perceptual coding

Lossy compression relies on predictable, redundant or psycho-acoustically non-salient information in audio to reduce its size for transmission or storage. The goal of lossy codecs is to maximize compression while maintaining acceptable fidelity to the uncompressed signal. Due to the significant redundancy and predictability in speech, lossy coding can often achieve dramatic reductions in audio size without compromising intelligibility. Since the goal of speech coders is to maintain comprehensibility of phonetic information it is unclear how much speaker information is removed or degraded.

2.1. Overview of Speex

This study bases its investigation of lossy compression on the recently developed Speex [3] [4] compression tools. There are 3 reasons why Speex was chosen as the basis for this study: 1) Speex is freely available, copyright, royalty and patent free and thus Speex can be easily incorporated into speaker recognition systems for potential mitigation of compression from other lossy compression techniques, 2) Speex has a computationally efficient software implementation that is very fast on modern hardware, since it is designed for real time communications, 3) Speex has enormous flexibility in quality and compression, allowing for a multiplicity of potential tradeoff options for its use in environments where SR may play a role. Speex allows for a range of bit rates, from 2 kbits/s to 44 kbits/s.

Speex bases its compression on CELP [6] [7] or Code Excited Linear Prediction. CELP is a tried and true speech coding approach first proposed in 1985. By 1991 there was a DOD standard established for very low bit rate communications based on CELP. CELP is based on analysis-by-synthesis techniques and was designed to improve upon earlier compression approaches based on simpler Linear Predictive Coding algorithms [8] and is now the most-used speech coding algorithm [3].

2.2. Other lossy compression algorithms used in this study

Two additional compression approaches were used in this study to determine if Speex compressed models could be used to advantage on other, non-CELP codecs. GSM (Global System for Mobile communications) 6.10 [9] is a RPE-LTP (Regular-Pulse Excitation Long-Term Predictor) based codec. Like Speex it was designed for speech and compresses audio based on prediction and correlations in the signal. The standard GSM 6.10 codec compresses audio at full rate to 12.2 kbit/s. GSM is the standard for the vast majority of cellular communications in the world and is optimized for real time compression.

The Vorbis codec [5] uses a third approach to lossy compression based on an implementation of the Modified Discrete Cosine Transform (MDCT) [10]. Unlike Speex and GSM 6.10, Vorbis was primarily designed for music compression and is in the same family as the mp3 and Advanced Audio Codec (AAC) compression schemes. This family of coders is largely based on psychoacoustic and perceptual principles to eliminate or heavily compress non-salient aspects of the signal for human perception. Due to the computational complexity of psychoacoustic compression Vorbis is generally not used for real time communications.

3. Database

The database used in this study consisted of 240 conversations each lasting approximately 120 seconds each. The conversations were recorded using a Samson C01U Condenser USB microphone in a low noise meeting room environment. A total of 120 speakers were collected and each speaker participated in 2 sessions of conversation. The conversational sessions were separated in time by several weeks to several months in order to perform realistic cross-session evaluations. All sessions were held face to face with the microphone positioned on the table facing one speaker. The data collected from this primary speaker was the only data used in these experiments and the secondary speaker's audio was removed from all files. Each session resulted in approximately 60 seconds of target speaker speech, sampled at 8000 Hz, 16-bit pcm.

Cross-session data separated by significant time was important to representing realistic speaker recognition results. Initial tests from same day sessions run with the original 120 conversations had a closed set accuracy of 100%. This fell to 89.2% when the 120 additional sessions collected weeks later were used as test data; this cross-session set was the configuration used as the basis of this study.

4. The speaker recognition evaluation system

The Gaussian Mixture Model (GMM) and Universal Background Model (UBM) approach, developed by Reynolds [9], is used as the basis for speaker recognition in this study. In our implementation the front-end feature processing consists of mel-weighted and delta cepstra generated from a frame size of 20ms with 50% overlap. During recognition, the likelihood of the test speech is computed for each of the GMMs produced during training. Only 5 mixtures are used for the calculation of the likelihood of a particular speaker’s GMM model, and the five mixtures are chosen from the most probable mixtures in the UBM. This study does not focus specifically on the accuracy of a given speaker recognition system in order to compare or improve algorithms, rather the goal is to demonstrate the effect of the speech coding conditions on a very common approach to speaker recognition.

5. Experiments

All experiments were conducted using the first session of the cross-session corpus described in section 3 for each speaker for training models and the second session for testing. Experiments evaluated three main conditions: the impact of coding on SR, conditioning models to mitigate compression mismatch and use of Speex on data compressed with other perceptual coders

All data was evaluated using a closed set/forced decision measure in which the top model returned in each test was counted as the winner. When the test file and the top model matched this was counted as correct. All other outcomes were counted as errors. This evaluation approach produced results that correlated highly with equal error rate measurements used in the speaker verification modality.

5.1. Impact of Speex coding in compression match and mismatch conditions

The first experiment examined the impact of Speex compression on speaker recognition at each quality level from 10 (least compressed) to 0 (most compressed). For this experiment each file from session 2 was compressed at each compression quality level from 10 to 0, plus the control set, which consisted of the uncompressed data. The Speex utilities *Speexenc* and *Speexdec* from Speex 1.2 were used for all compression and decompression. Once files were compressed they were converted back to 16-bit pcm files sampled at 8000 Hz with *Speexdec* for purposes of speaker recognition. Models were built from the full, uncompressed recordings from the first session and these models were then tested on all eleven of the test sets.

The second set of experiments evaluated the effect of matching test and train compression levels. This functioned to determine a reasonable level of Speex compression possible for use in storage and transmission of audio that may be used in speaker recognition technologies. In these experiments both the test (session 2) and train (session 1) sets were compressed to the same level, from 0 to 10, and matched evaluations were run.

5.2. Can SR models be conditioned to be resistant to compression mismatch?

Another research question of this study deals with model conditioning and whether Speex can be used to prepare a single set of models so that accuracy is maintained across multiple compression conditions. This would be useful in cases where the level of compression of the test data cannot be determined or where it may vary across a set of data. Since we have fast and easy access to real time Speex compression for model building the natural approach would be to evaluate combinations of Speex compressed train sets to determine an optimal set. Our initial experiments looked at the efficacy of training models at moderate compression levels (4-8) and testing across all conditions. As one might expect these models performed well on compressed data closest to the train compression level, but fell off quickly as the mismatch increased.

A second set of experiments combined sets of data compressed at different levels to see if this would provide robust performance across multiple compression quality levels. Much better results were obtained by combining train sets compressed at 4 and 8 levels together into a single model. Adding in the original (uncompressed) train set further improved accuracy on the high end.

5.3. Are Speex models effective on data coded with other lossy compression techniques (GSM, Vorbis)?

The final set of experiments examined the utility of Speex compression for improving SR models accuracy on other lossy compressed data. We chose to evaluate GSM and Vorbis, two very commonly used compression approaches that are based on different compression techniques than Speex. Since GSM has patent and other usage rights issues [3] and would be difficult to incorporate into an SR system for model conditioning, Speex is completely free of such issues and can be readily used for model enhancement.

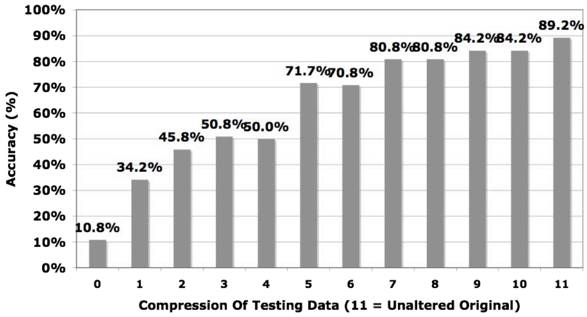
To test our hypothesis the test data (session 2) was compressed using both Vorbis and GSM 6.10 at the default settings in the *sox* speech conversion toolset. This data was then evaluated using SR models from all the compression quality levels generated from experiments described in 5.1 and the conditioned models from 5.2.

6. Results

6.1. Compression mismatch performance

The first experimental condition, testing compressed data with uncompressed models, demonstrates the sensitivity of SR to Speex compression. This is particularly true for quality levels of less than 9, where accuracy has dropped by almost 10% (table 1).

Table 1. *Effect of SPEEX Compressed Test Files On Closed Set Speaker ID (120 Cross-Session Speakers)*

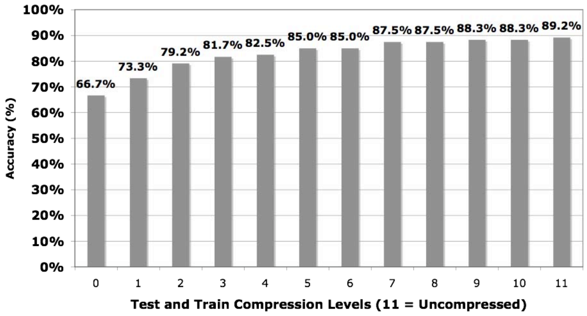


Indeed in mismatched conditions even a quality level of 10 incurs a significant loss in performance of 5%. At the highest compression/lowest quality level (0) performance is greatly impacted, and is only slightly above 10%.

6.2. Performance on matched compression levels

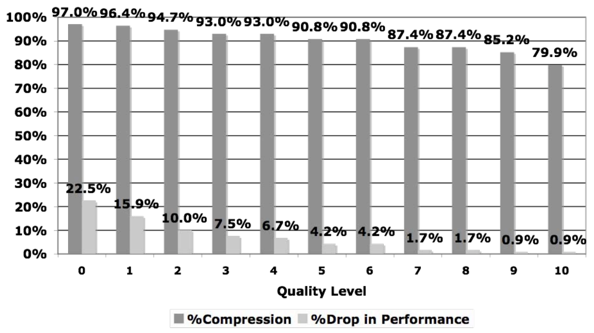
Results on matching compression levels (table 2) clearly show a dramatic improvement in performance across compression levels.

Table 2. *Accuracy of Matched Compression Levels for Test and Train*



A 10% drop in accuracy is not reached until compression level 2, and compression quality levels 9 and 10 have a reduction in accuracy of less than 1% when compared to the testing uncompressed data on uncompressed models. This result is very important in demonstrating the potential utility of Speex compression in the speaker identification environment. The trade-off between degree of compression and loss in accuracy in table 3 reveals that one can reduce the size of the audio used in speaker recognition environment by greater than 85% and suffer less than 1% drop in performance at compression quality 9.

Table 3. *Average reduction in file size for each quality level vs. impact on accuracy*



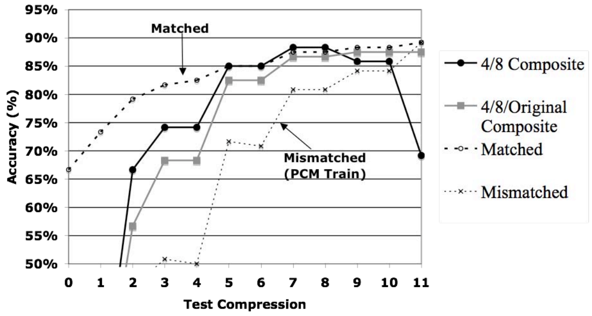
This makes Speex an impressive candidate for use in environments where moving or storing large audio files is problematic. In addition, this provides a useful means for storing samples of a speaker's voice with the model of that speaker for human verification purposes.

6.3. Model conditioning

The third experimental condition looks at the construction of a single model set that is resistant to the effect of lossy coding. As we saw in 6.1 a mismatch between coded data can be devastating. However, there may be cases where the degree of compression, or even whether data was compressed at all is an unknown quantity. The conditioning approach was to simply examine different combinations of combined compressed data in the training of speaker models. Many combinations were evaluated, but two proved to be particularly effective for a range of compression levels.

As one can see in table 4 a combination of 4 and 8 compressed data in a speaker model improved accuracy slightly over matched on quality levels 7 and 8 and equals matched accuracy on 5 and 6 and is within 2% for 9 and 10.

Table 4. *Composite Speex Models vs Matched and Mismatched Tests*

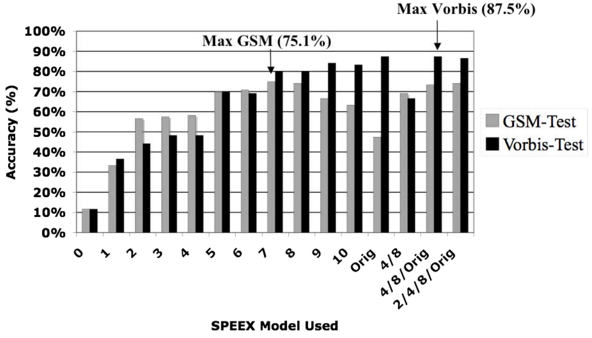


Adding in the original uncompressed data improves accuracy slightly at 9 and 10 over the 4/8 combination at the expense of a loss at 5 and 6, but high accuracy is maintained for uncompressed audio, where the 4/8 combined models perform poorly. In all cases the 4/8 combination and 4/8/uncompressed models outperform the model mismatch condition.

6.4. Speex SR models on other lossy compressed audio types

The use of Speex compression to mitigate the effects of other lossy compression approaches, namely GSM and Vorbis, is the final experimental condition. Results in table 5 show a significant difference between GSM and Vorbis coding. Using uncompressed pcm models for speaker recognition on GSM coded data has an accuracy of only 47%. Using any model compressed with Speex (other than extremely low quality compression levels of 0 and 1) improves this performance, with Speex compression level of 7 improving performance by absolute 28% to 75% correct recognition.

Table 5. *Accuracy of Speex-Models Tested on GSM and Vorbis Compressed Data*



Vorbis, being a psycho-acoustic approach used primarily for music, was not amenable to the use of Speex compression as a way of mitigating mismatch between test and train data. At the default settings for Vorbis used in these experiments only models containing the original uncompressed data perform best on Vorbis compressed data. The uncompressed data models dropped by only 1.7% from the matched results. Vorbis achieves compression rates of 78% on this data, but it is not a real time optimized speech coder like GSM or Speex. More importantly, with matched test/train at the 8 compression level Vorbis had no impact on SR performance. It is certainly promising to see data with such high compression rates performing so well with uncompressed models in SR experiments and even more impressive to have a music codec compress speech by 78% with no impact on SR.

7. Discussions/Conclusions

7.1. Utility of Speex for compression in speaker recognition audio

The results of the project demonstrate the potential utility of the Speex lossy compression system as a way of greatly reducing the size of audio files for use in the speaker recognition environment. In fact, in this study files were able to be reduced by an average of 85% while suffering less than 1% degradation in SR performance. This is a clearly beneficial tradeoff in scenarios where transmission and storage of full pcm audio presents a problem due to limitations on storage space, transmission bandwidth or transport of files. It is also plain from this study that high levels of compression can have a very large impact on performance, even when test and train data is compressed with the same quality levels. At the extreme end (0 quality) this resulted in a drop in accuracy of 22%.

7.2. Model conditioning

Compression mismatch had a significant negative impact on SR in these experiments. Test data at the 0 quality level tested with pcm models resulted in accuracy falling by absolute 70%. Even mismatch at 10 quality level harmed performance by 5%. To mitigate the effects of mismatch in scenarios where the degree of compression is unknown models were devised that mixed compression levels for a given speaker. This process generated a model set that was able to maintain matched performance over a large range of compression settings, and always out performed the mismatched cases.

7.3. Speex on other lossy coded data

Tests from this project were able to demonstrate that Speex models have a clear advantage in SR on GSM coded data over full bit-rate pcm models, with an improvement of 20% absolute accuracy. This is useful due to the complete lack of restrictions on using the Speex library and the copyright issues with GSM which would complicate model conditioning. Performance on Vorbis data, on the other hand, was not improved via Speex compression of models, and was best decoded with full pcm models.

7.4. Implications

This paper found that Speex is a viable option for speech compression in cases where SR will be used, as long as high quality levels are maintained. Further, Speex can be used to generate speaker models that are robust to varying compression levels and out perform pcm-based speaker models on GSM data. Additional research is needed to investigate the Vorbis coder, which, despite being designed for music, performed well as a compression technique useable with pcm-based speaker models and out-performed Speex with matched compression levels, albeit with a less drastic reduction in file size.

8. References

[1] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, and F. Pellandini, "Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition" EUSIPCO 2000, Tampere, Finland, 2000.

[2] M. Kuitert and L. Boves, “Speaker Verification with GSM Coded Telephone Speech”, Proc. Eurospeech’ 97, Vol. 2, pp. 975-978, 1997.

[3] Valin, J-M. "Speex: A Free codec for free speech", Australian National Linux Conference, Dunedin, New Zealand, 2006.

[4] Valin, J-M. *The Speex Codec Manual Version 1.2,* URL: http://Speex.org/docs/manual/Speex-manual.

[5] Xiph Foundation, *Vorbis I Specification*, URL: http://xiph.org/vorbis/doc/Vorbis_I_spec.html

[6] Schroeder, M. Atal, B. "Code-excited linear prediction (CELP): High-quality speech at very low bit rates", ICASSP 1985, Tampa, Florida, USA.

[7] Campbell, Joseph P. Jr., Thomas E. Tremain and Vanoy C. Welch, "The Federal Standard 1016 4800 bps CELP Voice Coder," Digital Signal Processing, Academic Press, 1991, Vol. 1, No. 3, p. 145-155.

[8] B.S. Atal, "The History of Linear Prediction," *IEEE Signal Processing Magazine, vol. 23, no. 2,* March 2006, pp. 154–161.

[9] Mouly, M and Pautet, M. *The GSM System for Mobile Communications*, Telecom Publishing, 1992.

[10] J. P. Princen, A. W. Johnson and A. B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," ICASSP 1987, Dallas, Texas, USA.

[11] Reynolds, D. A. “Comparison of Background Normalization Methods for Text-Independent Speaker Verification.” *Proceedings of the European Conference on Speech Communication and Technology,* Rhodes, Greece, 1997. Vol. 2, pp. 963-966.